

Weekly Report

2012.09.24-2012.09.30

张建霞

本周工作：

1. 论文阅读

- DICON：论文主要介绍了一种多维数据可视化的方法——dynamic icon-based visualization technique。方法以一种压缩图标 (compact glyphs) 形式显示多维数据的聚类 (clusters)。

- 设计原则：

- ◆ 聚类可视化的要点之一即每一个层级要有不同的粒度。DICON 采用的是 entity-feature-cluster 层次。
- ◆ 数据的不同维度和尺度应该采用同样的可视化编码方式。
- ◆ 相似的 cluster 的 icon 应该在视觉上相似，不相似的 cluster 的 icon 应该要很容易被区分。

- 每一个标签的设计都包括了如下的工作：

- ◆ 大小编码 (size encoding)

- 对于 n 维数据聚类的每一个实体 (entity)，都包括了一组特征 (features)，标记为 $F=f_0...f_n$ 。每个 feature 的值都被全局归一化到区间[0,1]。每个实体的所有 feature 值和为 1。这些 feature 值然后映射到颜色编码的小块 (cell)，cell 的 size 就代表了 feature 的值。多个实体被 pack 到一起组成一个 cluster，而 cluster 的标签大小和 cluster 内部的 entity 的数量成比例。

- ◆ 位置编码 (position encoding)

- 生成一个 cluster 的 icon 包括如下三步 :1)把每个实体的 icon 根据 feature 分割成独立的 cell , 2)把这些 cell 根据 feature 种类重新组合 , 3)把重组后的 cell 再合并成一个大的 icon。其中主要用到的是 treemap 的布局设计。

◆ 形状编码 (shape encoding)

- 采用了稳定的 (stabilized) Voronoi layout 方法 , 结合了 cluster 的数据分布的 kurtosis 和 skewness。Cluster 的 icon 采用的是梯形 , 上底的长度是 kurtosis , 峰值在水平方向的偏移是 skewness , 下底的长度随 cluster 实体数目的多少而动态调整。

◆ 颜色编码 (color encoding)

- 每种 feature 用不同的颜色 , 同时 , 颜色的饱和度 (saturation) 也用来编码。Saturation 编码有三种方案 , 分别可以可视化出 :1) cluster 的数据质量情况 , 2) 某些 feature 同时出现 (co-occurrence) 的情况 , 3) 某种独特的 (dominant) feature 的情况。

- 所有 icon 的分布采用了 global layout algorithm。根据不同的目的可以采用不同的方法 , 比如和地图结合 , 和 scatter plot 结合。DICON 同时也提供了基于 MDS 的投影方法 , 使 icon 基于它们的相似度进行分布。

- 动态变化 (animated transitions) , 变化的过程分为三个步骤 :1) 将 feature cell 拆分出来 , 2) 所有 feature cell 移动到目标位置并改变其形状 , 3) 所有的 feature cell 在新的组织方式下重组。步骤 2) 的实现依赖于 transition path bundling technique。每一个 feature cell 的运动路径由 4 个点 (此 cell 原来的质心→此 cell 原来所在 icon 的质心→此 cell 所在的新的 icon 的质心→此

cell 新位置的质心) 确定一条折线 , 根据这条折线确定的 B 样条曲线即为对应 cell 的运动路径。

- 用户交互 (user interaction) , DICON 系统支持了 merge、split、attribute grouping、filtering 和 highlight , 所有的这些方法都使得用户可以方便的进行数据交互式探索 , 修缮可视化结果。
- DICON 的长处
 - ◆ 编码了统计信息 , 促进了 cluster 的评估和调整
 - ◆ 支持大量的 cluster
 - ◆ 聚类质量评估、交互调整和探索
 - ◆ 新式的布局方法
 - ◆ 平滑的动态显示

2. 代码阅读

- 主要阅读了芯芯姐写的淘宝标签代码 , 了解了其现在所在的项目及其工作的进度 , 争取尽快融入项目组。

下周计划 :

1. 数据整理 , 将几个 txt 里面的信息合并到一个 txt 以便下一步使用。
2. 自学点数据挖掘方面的东西。